Evolutionary conservation of the polyproline II conformation surrounding intrinsically disordered phosphorylation sites

W. Austin Elam,[1] Travis P. Schrank,[2] Andrew J. Campagnolo,[1] and Vincent J. Hilser[1,3]*

[1]T.C. Jenkins Department of Biophysics, Johns Hopkins University, Baltimore, Maryland
[2]Department of Biochemistry and Molecular Biology, University of Texas-Medical Branch, Galveston, Texas
[3]Department of Biology, Johns Hopkins University, Baltimore, Maryland

Received 21 November 2012; Revised 7 January 2013; Accepted 9 January 2013
DOI: 10.1002/pro.2217
Published online 23 January 2013 proteinscience.org

Abstract: Intrinsically disordered (ID) proteins function in the absence of a unique stable structure and appear to challenge the classic structure-function paradigm. The extent to which ID proteins take advantage of subtle conformational biases to perform functions, and whether signals for such mechanism can be identified in proteome-wide studies is not well understood. Of particular interest is the polyproline II (PII) conformation, suggested to be highly populated in unfolded proteins. We experimentally determine a complete calorimetric propensity scale for the PII conformation. Projection of the scale into representative eukaryotic proteomes reveals significant PII bias in regions coding for ID proteins. Importantly, enrichment of PII in ID proteins, or protein segments, is also captured by other PII scales, indicating that this enrichment is robustly encoded and universally detectable regardless of the method of PII propensity determination. Gene ontology (GO) terms obtained using our PII scale and other scales demonstrate a consensus for molecular functions performed by high PII proteins across the proteome. Perhaps the most striking result of the GO analysis is conserved enrichment ($P < 10^{-8}$) of phosphorylation sites in high PII regions found by all PII scales. Subsequent conformational analysis reveals a phosphorylation-dependent modulation of PII, suggestive of a conserved "tunability" within these regions. In summary, the application of an experimentally determined polyproline II (PII) propensity scale to proteome-wide sequence analysis and gene ontology reveals an enrichment of PII bias near disordered phosphorylation sites that is conserved throughout eukaryotes.

Keywords: intrinsically disordered; phosphorylation; polyproline II; gene ontology; proteome

Introduction

Nearly 30% of the human genome encodes intrinsically disordered (ID) protein sequences that assume no unique, stable structure under native conditions.[1] The prevalence of ID segments presents a challenge to the structure–function paradigm; because ID sequences lack stable structure yet perform crucial functions. ID proteins differ in amino acid composition from structured proteins,[2] resulting in higher charge to hydrophobicity ratios.[3] Presently, whether

Abbreviations: CD, circular dichroism; GO, gene ontology; ID, intrinsically disordered; ITC, isothermal titration calorimetry; PII, polyproline II; SH3, Src-homology 3; Sos, son of sevenless.

Additional Supporting Information may be found in the online version of this article.

Grant sponsor: National Science Foundation; Grant number: MCB-0446050.

**Correspondence to:* Vincent J. Hilser, Departments of Biology and Biophysics, Johns Hopkins University, 3400 N. Charles St., 107 Mudd Hall, Baltimore, MD 21218. E-mail: hilser@jhu.edu

or how compositional differences manifest as conformational propensities that may be tunable and thus utilized for functions such as signaling is not well understood. Recent studies examined ID protein sequences to identify molecular recognition features proposed to be local sequence regions prone to adopt structures important for protein-protein interactions.[4,5]

One conformation suggested to be highly populated in the unfolded states of proteins[6,7] and peptide sequences with high net charge[8–11] is the polyproline II (PII) conformation. Binding of proline-rich regions that are biased to the PII conformation is vital for cell signaling related to growth and differentiation.[12,13] Presently, it is estimated that the human genome may encode over 500 copies of domains (including SH3, SH2, WW, EVH1, and GYF) that interact with proline-rich regions.[14,15] In fact, genomic analysis has identified proline-rich regions as one of the most commonly encoded motifs in eukaryotes.[16] However, concerning the PII propensity of all amino acids (not just proline-rich regions) at a proteome-wide scale, little is known of the distribution of PII bias among protein sequences, the evolutionary conservation of PII bias within protein sequences, or the possible utilization of PII for functions outside of cell signaling.

To address whether sequences select for regions of high PII propensity and potentially utilize PII propensity functionally, a complete, calorimetrically-determined amino acid propensity scale is developed for the PII conformation. Amino acid PII propensity has been the focus of several experimental[17–21] and computational[22–25] studies. We employ our calorimetrically-determined PII scale and scales of others[17–25] to investigate whether PII is enriched in ID proteins, and if so what functional roles such proteins play. Our approach maps amino acid PII propensities onto protein sequences in order to characterize the functional roles of PII at a proteome-wide level.

Results and Discussion

PII propensities for all amino acids were determined using a peptide host-guest system and isothermal titration calorimetry.[26–28] The *C. elegans* Sem-5 (sex muscle five) SH3 (Src-homology 3) domain binds a peptide corresponding to the recognition sequence of its binding partner, Sos (son of sevenless). Importantly, in the Sem-5 SH3-Sos complex, the Sos ligand is bound in the PII conformation[29] [Fig. 1(A)]. Substitution at a noninteracting position results in a decrease in binding affinity (K_{app}) relative to the wild-type peptide. Comparison of K_{app} from isotherms [Supporting Information Fig. 1(A)] reports on the change in binding affinity upon substitution. Because the substitution is made at a site that is surface exposed in the bound complex and does not perturb the binding interface [Supporting Information Fig. 1(B)],[26,29] we can infer that the observed change in binding affinity reflects a change in the conformational equilibrium (K_{conf}) between binding incompetent and binding-competent (i.e., PII) states of the Sos ligand [Fig. 1(A)]. After correction for effects of the guest residue on *cis*/*trans* isomerization of the preceding proline residue using ^{1}H-NMR experiments,[30,31] the measured change in K_{app} can be interpreted as a PII propensity[26] (Supporting Information). Cyclic sidechains (HIS, PHE, TRP, TYR) and GLY have low PII propensities, while long, charged sidechains (GLN, GLU, LYS) have higher PII propensities [Fig. 1(B)]. However, there is no apparent correlation of PII propensity with any single physico-chemical property.[32] Being charged, for example, does not necessarily correspond to high PII propensity, as ASN, ASP, and HIS are average or below [Fig. 1(B)].

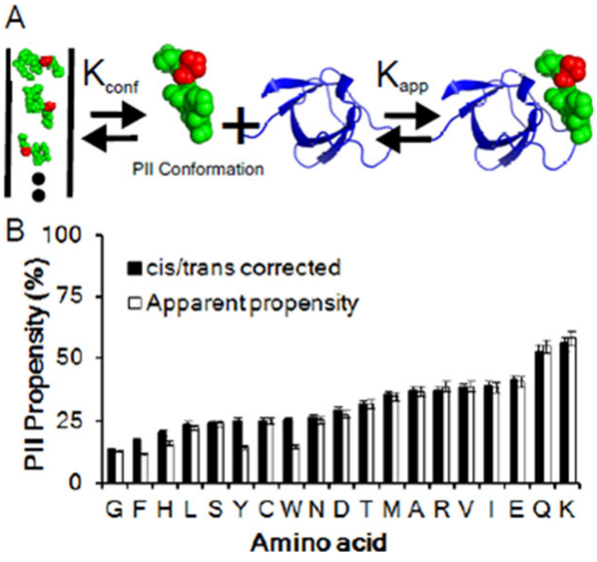


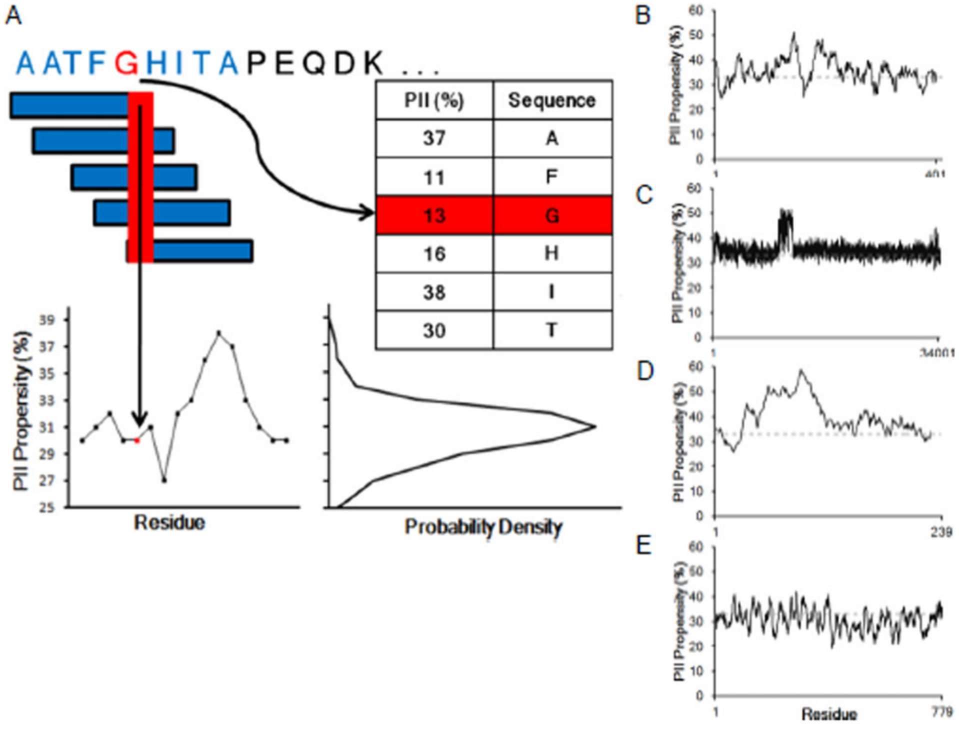
Figure 1. Determination of a calorimetric PII scale. (A) Schematic of the binding equilibrium of SH3 (blue) and Sos peptide (green) with surface-exposed substitution site (red) (PDB code: 1SEM).[29] (B) PII propensity scale with *cis/trans* isomerization correction (black), and without correction (white). Error bars are propagated error in $\Delta\Delta G$ + 30 cal/mol. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Differences in free energy of binding calculated for ALA and GLY correspond to previous measurements,[26] and agree with PII propensities determined by others.[22,33] A study using a hard sphere collision model recently evaluated the ability of conformational bias to different regions of Ramachandran space to reproduce experimentally measured binding of ALA and GLY substituted Sos peptide to SH3.[34] The results of the simulations indicated that only bias to the PII conformation at rates similar to those measured previously[26] and in this study were capable of quantitatively reproducing experimentally determined binding energies.[34] In fact, the same hard sphere model used by Whitten *et al.*[34] can quantitatively reproduce the PII propensities for ALA and GLY relative to proline.[32] Further, the rank order of PII propensities determined by our calorimetric scheme can be reproduced by CD spectroscopy of representative Sos peptides [Supporting Information Fig. 1(C)]. Together, these data support the validity of our calorimetrically determined PII propensity scale.

The experimental PII propensity scale reported here [Fig. 1(B) and Supporting Information Table I] was used to investigate the PII content of representative proteomes. An algorithm was developed to calculate PII propensity along a given sequence by determining the average PII bias within a sliding window [Fig. 2(A)]. The effectiveness of this approach was determined by examining sequences previously reported to be high in PII [Fig. 2(B–E)]. The high PII regions of human tau,[35] the PEVK domain of the enormous human titin protein,[36] and the periplasm-spanning domain of the bacterial TonB protein,[37] are all detected by our algorithm as being significantly above the average PII propensity. In contrast, the PII propensity calculated along the sequence of an outer membrane protein, known to have a β-barrel structure,[38] appears as noise about the average PII propensity [Fig. 2(E)]. The ability of the algorithm to (1) detect high PII regions in systems investigated by others and (2) discriminate between these regions and proteins known not to have significant PII structure, suggests that our algorithm can reasonably detect the level of PII bias in protein sequences.

The distributions of average PII propensities for structured protein segments (extracted from the PDB[39]) and ID regions from DisProt[40] (which have been experimentally verified to be disordered or contain disordered segments) reveals important differences (Fig. 3). Although there is considerable overlap in the distributions, ID regions show enrichment of high PII propensity. Specifically, 92% of all windows with an average PII propensity of 40% or more occur in ID regions. All of the most extreme PII segments (i.e., >47% PII) are in disordered sequences. Importantly, this is not to claim that all ID segments are high in PII bias. Notably, ID sequences often contain GLY-rich regions that are on the low end of the PII distribution.

To determine whether the computed enrichment of PII in ID sequences is dominated by a small number of residues, or whether a broad repertoire of amino acids contribute to the signal, PII propensities were recomputed with PRO, LYS, and GLN artificially set to mean values [Supporting Information



PII (%)	Sequence
37	A
11	F
13	G
16	H
38	I
30	T

Figure 2. Schematic and validation of an algorithm for calculation of PII propensity within amino acid sequences. (A) A sliding window calculated the position-specific average PII propensity along a dataset of protein sequences, referencing the experimentally determined propensity at site (red) and window residues (blue). High PII regions of (B) human tau, (C) human titin, and (D) bacterial TonB, can be detected using the sliding window scheme and the calorimetrically-determined PII scale (black line). The proteome average sequence PII propensity is shown (dashed gray line) for reference. (E) A transmembrane protein (Omp85) whose family is known to have β-barrel structure[38] shows no high PII signal. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

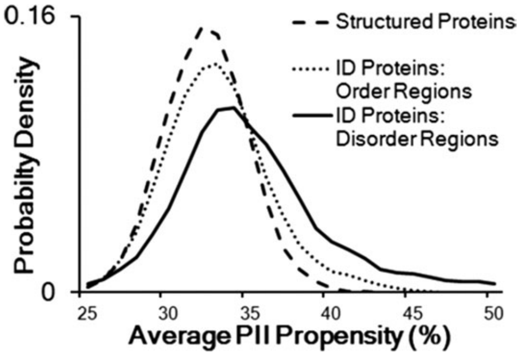


Figure 3. PII propensity is enriched in ID segments. ID segments show significant ($P < 0.05$) enrichment of high PII propensity (solid line) relative to structured proteins (dashed line). Ordered segments of disordered proteins (dotted line) are shown.

Fig. 2(A)]. ID segments nonetheless maintained their relative enrichment, indicating that enhancement of PII content is robustly encoded and not an artifact of selecting only PRO, LYS, or GLN rich segments. The observed enrichment was also insensitive to the window size employed to calculate the PII bias [Supporting Information Fig. 2(B,C)]. The enrichment of PII in ID sequences was also observed with PII propensities measured by others[17,19,22] [Fig. 4(A–C)]. However, the enrichment could not be captured by randomly generated PII scales [Supporting Information Fig. 2(D)]. In summary, the fact that the enrichment observed in Figure 3 could also be captured using other experimental or computational (coil library or molecular dynamics) scales, indicates that the enrichment is strongly encoded and not dependent on the method of PII determination.

We note that Avbelj and Baldwin report that neighboring β-branched or aromatic residues may promote β-strand conformations[41] in a coil library. Another study reports that aromatic amino acids may disfavor the PII conformation.[21] Our calorimetrically-determined PII propensities are consistent with these data,[21,41] as most cyclic amino acids in our scale have low PII propensities, even with *cis*/*trans* isomerization corrections [Fig. 1(B) and Supporting Information Table I]. Differences in amino acid composition between our datasets [Fig. 4(D)] directly support our expectation that aromatic residues contribute only modestly to our analysis.

Importantly, in our sequence analysis, we assume that context dependence of PII propensity is minimal, which is supported by recent studies of blocked dipeptides.[20] Correlation of PII propensities reported by Pappu and coworkers[22] also suggests that nearest neighbor context has negligible impact on the rank order of PII propensities (Supporting Information Table II). PII propensities calculated in PRO, ALA, GLY, VAL, and PHE contexts are all statistically correlated ($P < 0.05$), with Spearman coefficients ranging from 0.800 to 0.979 (Supporting Information Table II). While the numerical PII propensities differ in these contexts,[22] it is clear that

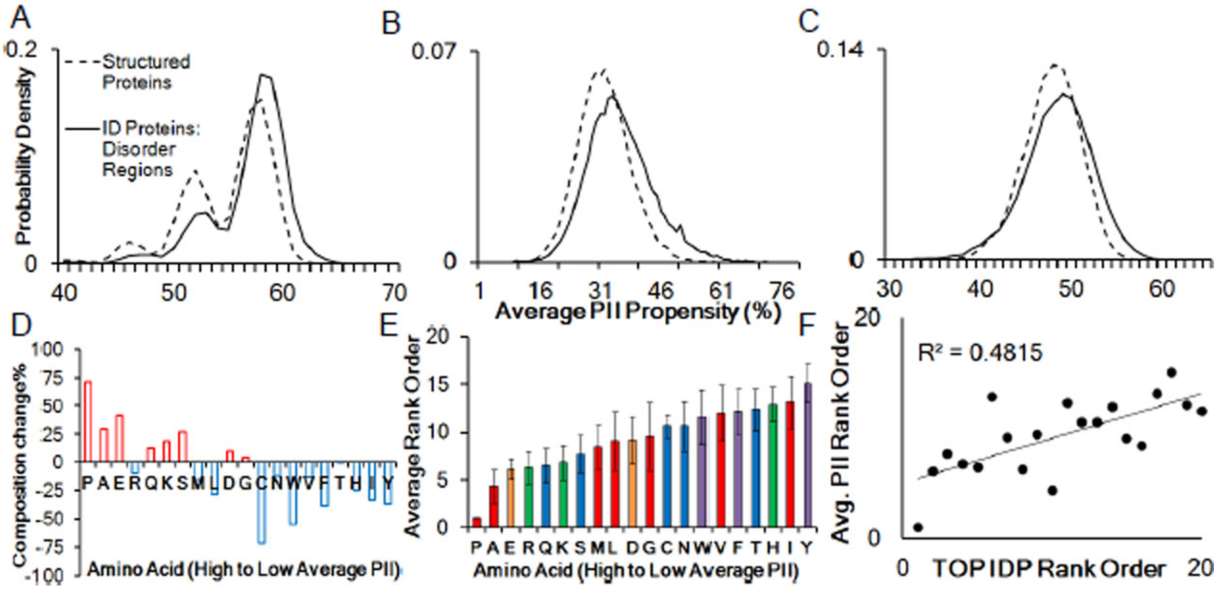


Figure 4. Enrichment of PII propensity in ID sequences is detected using other PII scales. ID segments of disordered proteins show significant ($P < 0.05$) enrichment of high PII propensity (black line) relative to structured proteins (dashed line) using PII scales of (A) Rucker *et al.*,[17] (B) Tran *et al*. PPXPP,[22] and (C) Grdadolnik *et al.*[19] (D) Amino acids whose frequencies increase in ID sequences (red) are near the top of the average PII rank order; those with decreasing frequencies in ID sequences (blue) occur near the bottom of most PII scales. (E) Average PII rank order from all PII scales.[17–25] Error bars show standard deviation of PII rank across scales. Amino acids are nonpolar (red), polar (blue), aromatic (purple), negatively charged (orange), positively charged (green). (F) Spearman correlation of the average PII rank order to the TOP-IDP scale. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.][43]

host context does not significantly change the rank order of a PII scale (Supporting Information Table II). These observations support our assertion that low and high PII sequences can be differentiated in the proteome. Most importantly, PII scales derived from host systems of different contexts including biological peptides (this study), proline-rich peptides,[17,22] and dipeptides[19] all detect the enrichment of PII bias within ID proteins [Figs. 3 and 4(A–C)], suggesting the coding of PII in these regions is robust despite possible near neighbor effects.[41,42]

Analysis of the amino acid composition of sequence datasets elucidates the ability of multiple PII scales to detect the enrichment of PII in ID segments because, even though the numerical amino acid PII propensities may differ between scales, there is a general consensus regarding which amino acids have high (PRO, LYS, GLN, GLU) and low (HIS, TRP, TYR, PHE) PII propensity. The consensus can be clarified by examining the average rank order of all PII scales, which shows that amino acids with high average PII rank from all scales also tend to be enriched in amino acid composition within ID sequences [Fig. 4(E)]. Further, the PII rank order averaged from all scales correlates ($P < 0.05$) with the rank order of the TOP-IDP scale, a scale of amino acids proposed to promote disorder[43] [Fig. 4(F)], lending further evidence to the hidden correlation between PII scales and their ability to detect enrichment in ID segments.

To investigate whether PII distributions in ID and structured proteins arise from enrichment within specific local sequence stretches, the distributions were compared to artificial sequences constructed by shuffling sequences within each group. PII propensities of shuffled sequences of structured proteins showed no change from the original distribution [Fig. 5(A)], indicating that PII bias is randomly encoded (i.e., not enriched) along sequences of structured proteins, a result consistent with proteome-wide sequence correlations reporting nearly random site-to-site correlations between amino acids in sequences of structured proteins.[44] Shuffling of ID sequences, in contrast, reveal a dramatic change ($P < 0.05$) in the distribution [Fig. 5(B)], suggesting that PII is selectively enriched within particular segments of the ID sequence.

To determine the extent to which evolution selected for high PII sequences, *in silico* evolution was performed to monitor robustness of sequence PII propensities to amino acid substitution. Substitutions were performed by two methods: (1) substituting randomly, but maintaining the dataset amino acid composition, and (2) using the well-established BLOSUM62 matrix,[45] which generally preserves physico-chemical properties of substituted amino acids. Regardless of substitution method, the PII propensities of structured protein sequences were maintained [Fig. 6(A)], further supporting the conclusion from sequence shuffling [Fig. 5(A)] that PII in structured proteins is not an evolutionarily selected trait.

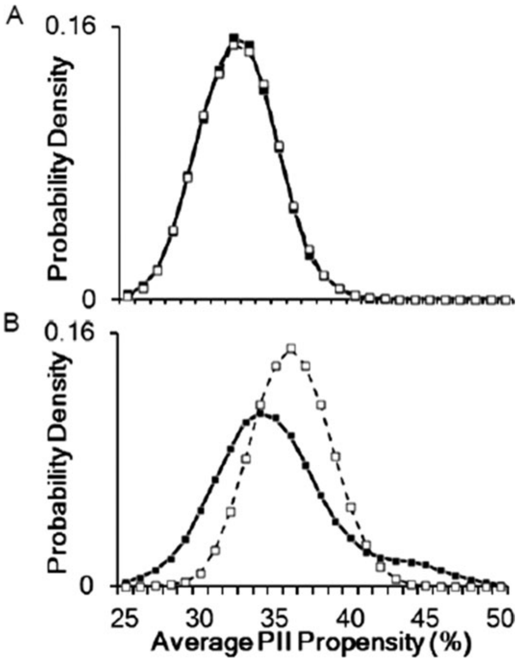


Figure 5. (A) The PII propensity distribution of folding sequences (■) and shuffled sequences (□) are superimposable. (B) Locally enriched high PII bias (■) is abolished upon shuffling ID sequences (□).

Unlike the sequences of structured proteins, substitution of ID sequences, either randomly or by BLOSUM62, resulted in a significant decrease in the PII propensity [Fig. 6(B,C)]. These results are consistent with bioinformatics analyses suggesting that, relative to sequences encoding structured proteins, conservation of ID within a sequence is more difficult in *in silico* evolution experiments.[46] Sensitivity to substitution, even when the physico-chemical properties are maintained, indicates that the high PII segments occupy a small, highly specialized sequence space that has evolved specifically to preserve PII propensity.

What function do these specialized, high PII proteins perform? To address this question in a global, systematic, and unbiased way, the PII content of six eukaryotic proteomes was calculated [Fig. 7(A)], and gene ontology (GO) analysis was performed for the top 1% of PII proteins from each proteome. The proteins in the top 1% of each eukaryotic proteome exhibited strong conservation of features and functions. Of the top GO terms returned in order of statistical enrichment, many were identical in all proteomes [Fig. 7(B) and Supporting Information Table III]. High PII proteins are associated with a diverse array of functions (Supporting Information Table III), not for one specialized purpose. One GO term of note that was reproducibly enriched was

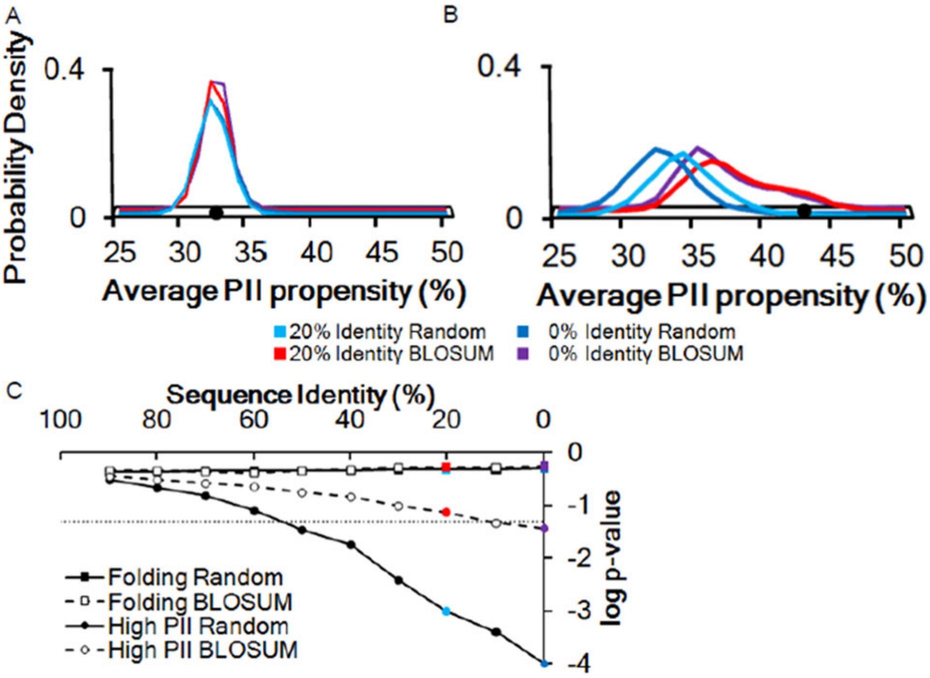


Figure 6. Evolutionary selection of high conformational bias. (A) Amino acid substitution by either random change or by BLOSUM62[45] preserves the average PII propensities of 10,000 *in silico* evolved "daughter" sequences generated from a "parent" set of randomly selected folding protein sequences ($P > 0.25$). The average PII propensity of the "parent" sequences (black dot) is well within the "daughter" distribution. (B) The average PII distributions of 10,000 "daughter" sequences deviate from the high PII "parent" sequences (●) ($P < 0.05$). (C) The statistical deviation of the "daughter" sequence distribution depends on the sequence identity maintained during *in silico* evolution. The "parent" sequence falls within the "daughter" PII distribution at all levels of sequence identity in structured proteins (squares), regardless of whether substitution was performed randomly (■) or by BLOSUM62[45] (□). The log of the p-value of the "parent" sequence PII propensity relative to the "daughter" sequences sharply decreases in high PII proteins when substitution was random (●) and when substitution occurred by BLOSUM62[45] (○) Colored points on each line correspond to the significance of the "parent" to "daughter" difference shown in (A,B). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

"collagen," an archetype PII triple helix.[47] In addition, transcription regulation also appears to employ high PII proteins, consistent with the observation that transcription factors are enriched in ID.[48] Similar GO terms were obtained with other PII propensity scales that utilize experimental or computational (coil library and molecular dynamics)[17–25] methods [Fig. 7(C)]. A consensus on the GO features of high PII proteins was evident from comparison of results obtained using different PII scales; in stark contrast to that observed using randomly selected protein sets.

A striking feature of the analysis is that high PII proteins have a remarkable propensity for phosphorylation ($P < 10^{-8}$) [Fig. 7(B)], adding clarity to the comparatively weak correlation observed between phosphorylation sites and disorder.[49] Statistical enrichment of the "phosphoprotein" GO term was robust, being observed independent of including PRO, SER, THR, or TYR in the calculation of the top PII proteins in the proteome [Fig. 7(D)]. Enrichment of the "phosphoprotein" GO term among the highest PII sequences in the proteome was also detectable using other PII scales[17–25] [Fig. 7(D)]. Calorimetric determination of the impact of phosphorylation of SER, THR, and TYR [Fig. 8(A)] revealed amino acid specific effects. While the PII propensity of SER and TYR are not affected by phosphorylation (within error), a dramatic increase in the PII propensity of phospho-THR was observed, reaching a value that compares to the high PII seen for PRO residues.

Investigation of the origin of this effect reveals that the steric consequences of phosphorylation at THR are significantly higher than with SER, a result that is not altogether unexpected given that THR phosphorylation introduces additional bulkiness to the β-carbon (Supporting Information Fig. 3). Such changes are qualitatively similar to mutational strategies that modulate accessible conformations by introduction or removal of β-branched amino acids.[50] Phosphorylation of SER or TYR, on the other hand, produces no such effect. We note that in the context of the Sos peptide system utilized here, SER, THR, and TYR are observed to behave differently upon phosphorylation. In other contexts, however, phosphorylation or other post-translational modification may have other effects on conformational propensity.

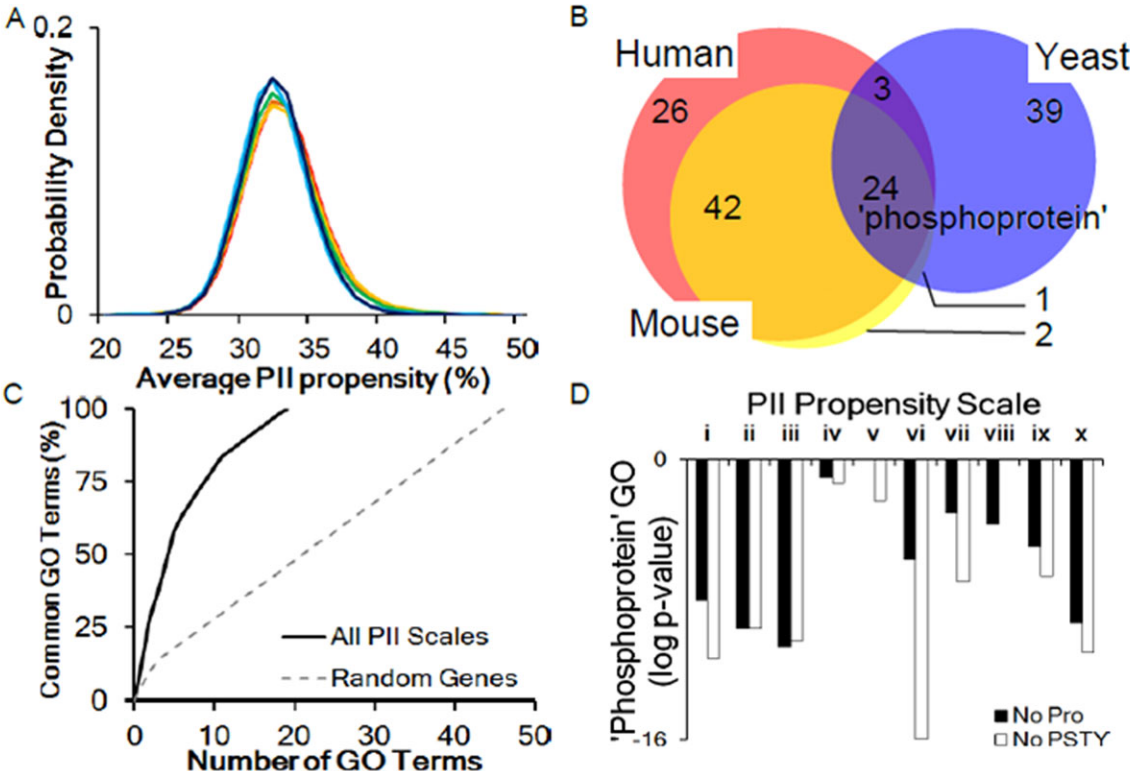


Figure 7. Phosphorylation is a functionally conserved feature of high PII proteins. (A) Average PII propensity distributions for six eukaryote proteomes: *H. sapiens* (red), *M. musculus* (orange), *D. melanogaster* (yellow), *C. elegans* (green), *A. thaliana* (light blue), and *S. cerevisiae* (dark blue). (B) Venn diagram of the number of features shared in the top 1% of PII proteins from *H. sapiens* (red), *M. musculus* (orange), and *S. cerevisiae* (blue). (C) Commonality of the top five GO terms by statistical enrichment among PII scales[17–25] and the present scale (black line) compared to ten random protein sets (gray dashed line). (D) "Phosphoprotein" GO term enrichment obtained by PII scales: (i) Rucker *et al.*,[17] (ii) Shi *et al.*,[18] (iii) Grdadolnik *et al.*,[19] (iv) Oh *et al.*,[20] (v) Brown *et al.*,[21] (vi) Tran *et al.* (PPXPP),[22] (vii) Fleming *et al.* (coil library),[23] (viii) Beck *et al.*,[24] (ix) Moradi *et al.*,[25] and (x) the present scale. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

In any case, our results [Fig. 8(A)] indicate that post-translational modifications have a capacity to dramatically change local bias toward the PII conformation.[51] Biologically, this result provides a compelling mechanism for local "tuning" of backbone conformational bias, which may be exploited for multiple functions including molecular recognition, targeting for degradation, or allosteric regulation.

Although gene ontology reveals that high PII proteins are often phosphorylated, it is not clear whether these proteins are phosphorylated within high PII regions or at other locations in the protein. Analysis of the PII propensities of known, experimentally validated phosphorylation sites[52,53] indicates that phosphorylation sites are indeed coincident with higher PII segments [Fig. 8(B)]. The contexts of phosphorylation sites are diverse, consisting of regions composed of charged residues as well as PRO-rich regions containing few charged residues. Analysis of phosphorylation site density as a function of PII propensity reveals a striking enrichment of phosphorylation sites in both high and low PII contexts [Fig. 8(C)]. The increased density at high PII is surprising given that the PII propensities of phosphorylation-competent residues (SER, THR, and TYR) have average or low PII propensity [Fig. 1(B)]. Of note is that the enrichment in phosphorylation site density is common throughout many eukaryotic proteomes, including human [Fig. 8(C)], mouse, fly, and yeast (Supporting Information Fig. 4). Because the majority of phosphorylation sites occur at SER, these sites dominate the enhancement distributions [Fig. 8(C) and Supporting Information Fig. 4]. Inspection of THR phosphorylation site density, however, shows a distinct preference for localization in high PII contexts. The enrichment of THR in the high PII contexts, which is the only residue to show a phosphorylation dependent increase in PII propensity (i.e., tunability) in our host guest system, strongly suggests that nature has specifically selected THR for tuning of local conformational bias in high PII regions. It is remarkable that THR exhibits alternate distribution relative to SER, particularly in light of a recent study suggesting THR sites evolve at different rates than SER or TYR sites.[54]

The distributions of site densities observed in Figure 8(C) (also observed in other eukaryotes, Supporting Information Fig. 4) prompted investigation of the correlations of PII propensity with phosphoprotein functions at a proteome-wide level. Figure 9

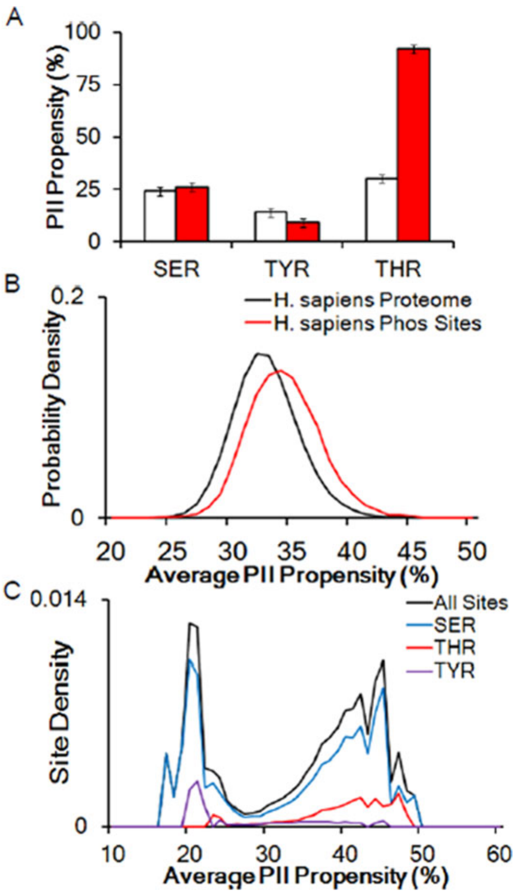


Figure 8. Phosphorylation can modulate PII and is differentially distributed in the proteome. (A) Calorimetrically-determined PII propensities for unmodified (white) and phosphorylated (red) SER, THR, and TYR. (B) PII propensity distribution of the *H. sapiens* proteome (black) and phosphorylation sites therein (red). (C) Bimodal enrichment of phosphorylation site density (black) observed in *H. sapiens*, dominated by SER (blue). THR sites (red) have enriched density in high PII contexts, while TYR (purple) is mostly in low contexts. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

reports a condensed representation of the different biological processes and molecular functions associated with phosphoproteins whose experimentally validated phosphorylation sites reside in either low or high PII contexts. The low PII phosphoproteins are typically proteins involved in kinase or transferase activity [Fig. 9(A,B)]. These proteins employ SER, THR, and TYR phosphorylation sites. The relative enrichment of TYR sites is expected from the site density [Fig. 8(C)], and the identities of the low PII phosphoproteins (kinases and transferases) is consistent with lower PII sequences tending to have globular structures [Figs. 3 and 8(B)]. TYR enrichment is also in agreement with amino acid composition bias we have noted for structured proteins [Fig. 4(D)]. In contrast, the high PII phosphoproteins rarely employ TYR sites in the human proteome [Figs. 8(C) and 9(C,D)]. Close examination of Figure 9(C,D) reveals that SER and THR sites exhibited very similar GO term enrichment, with "macromolecular complex assembly" [Fig. 9(C)] being the only GO term exclusively enriched in high PII THR sites. High PII phosphoproteins are enriched in a biological processes and molecular functions associated with mitosis, chromosome organization, cell cycle regulation, and transcription (consistent with previous results suggesting transcription factors are prone to be ID[48]) [Fig. 9(C,D)]. Comparing the difference in amino acid utilization (color) and the GO terms listed in Figure 9, it is immediately evident that low and high PII phosphoproteins perform different cellular functions. Perhaps even more interesting, we note that the GO terms enriched for high PII phosphoproteins [Fig. 9(C,D)] differ slightly from those observed for all high PII proteins from the proteome at-large (Fig. 8 and Supporting Information Table III), suggesting that modification (phosphorylation, in this case) within high PII context may be selected for a specialized functional utility.

Conclusions

Here we demonstrate for the first time a proteome-wide correlation between an experimentally determined conformational bias for PII and the propensity to be intrinsically disordered and thus unfolded. The functional importance of this relationship is revealed through a dramatic enrichment of phosphorylation sites within high PII segments. Proline-directed phosphorylation sites contribute to the enrichment of phosphorylation sites within high PII segments. Yet, there are hundreds of experimentally validated phosphorylation sites that also contribute to the enrichment but contain no nearby proline residues. We speculate that the proteome-wide bias of phosphorylation sites to high PII (and therefore likely disordered) segments may be a result of evolutionary pressures to facilitate kinase accessibility to these disordered regions. The conformational biases within these disordered regions may be a thusfar unappreciated means of regulating kinases or phosphatase accessibility and as a consequence their activity in signaling or other functions.

The conservation of these trends across multiple proteomes and the differential sensitivity of THR and SER to phosphorylation provide a compelling argument for their differential usage. We have endeavored to elucidate the different biological processes and molecular functions for which THR and SER phosphorylation sites have been selected for, finding that the functions of these sites in low and high PII contexts are completely different. Not explored, but equally as plausible is the possibility that other post-translational modifications (acetlyation, methylation, etc.) may also utilize ID segments and perhaps even differentially impact PII conformational propensity as is the case with phosphorylation.

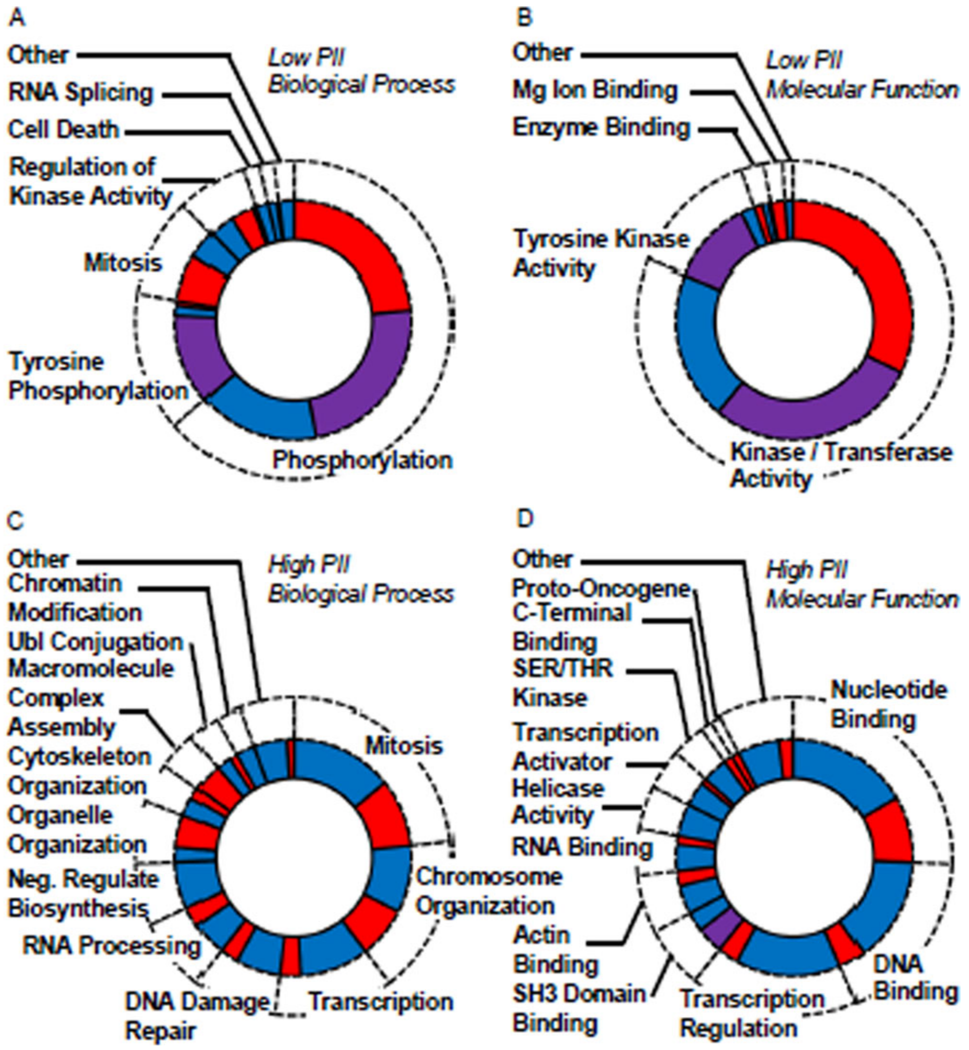


Figure 9. Low and high PII phosphoproteins are utilized for different cellular functions. Relative enrichment for GO terms obtained for SER (blue), THR (red), and TYR (purple) phosphoproteins correspond to the pie sizes in the above graphs. (A) Low PII phosphoprotein biological process related GO terms. (B) Low PII phosphoprotein biological process related GO terms. (C) High PII phosphoprotein biological process related GO terms. (D) High PII phosphoprotein biological process related GO terms. Comparison of the GO terms and amino acid utilization (SER, THR, or TYR) in each panel immediately shows how low PII (A,B) and high PII (C,D) phosphoproteins have different functions across the human proteome. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Our results reveal a potentially new way that ID proteins can modulate activity. Instead of using post-translational modification to induce a conformational change between two ostensibly discrete conformations (i.e., T and R states) in the context of a folded protein, ID proteins can potentially affect functional changes by tuning the distributions of otherwise disordered states specifically with respect to the PII conformation, as in the case for THR in this study. Whether and how this can be functionally utilized awaits further investigation.

Materials and Methods

ITC and CD experiments

The Sos peptide (Ac-VPPPVPPRRRY) and variants of the peptide with guest "X" at position (Ac-VPPX-VPPRRRY) were acquired commercially from GenScript USA or Neo BioSci. Purity of the peptide samples (>98%) were estimated using reverse high performance liquid chromatography and by mass spectrometry. Sem-5 C-terminal SH3 domain from *C. elegans* was purified as described previously.[29] The SH3 used in this study is a pseudo-wild type protein. CYS 55 has been mutated to an ALA to prevent possible oxidation and intermolecular cross-linking.

A Microcal VP-ITC system was used to perform all titration experiments.[55] SH3 was dialyzed against phosphate buffer, pH 7.5 (20 m*M* sodium phosphate, 200 m*M* sodium chloride (Fisher)). Lyophilized peptides were dissolved in buffer from the final protein dialysis. Protein and peptide sample concentrations were determined using the Edelhoch method.[56] Protein samples for titration experiments ranged in concentration from 0.5 to 0.65 m*M*, and peptide concentrations were approximately 10 times the protein concentration. All Sos peptide variants exhibited similar solubility. At 25°C, a series of 8 uL injections were made (34–35 total injections), with a spacing of 280 s between injections for equilibration. An initial injection of 2 uL was made and the data discarded for each titration to account for heat anomalies caused by instrument equilibration and pre-titration mixing by diffusion. Data were corrected for ligand heat of dilution by performing a titration of peptide into buffer and directly subtracting the resulting heats. These corrected data were fit in Origin 7 (OriginLab) using a nonlinear least squares regression varying the stoichiometry (n), binding constant (K_{app}), and the molar heat of binding (ΔH). The apparent free energy of binding (ΔG_{app}) and the entropy (ΔS) were calculated using the best-fit binding constant (K_{app}) and the thermodynamic relation below:$-RT \ln K_{\mathrm{app}} = \Delta G_{\mathrm{app}} = \Delta H - T\Delta S$(1)where R is the gas constant (1.985 cal/K/mol) and T is the temperature (298 K). PII propensities were determined from the $\Delta\Delta G_{\mathrm{app}}$ of each mutant relative to wild type (proline) as described previously.[26–28] Statistical analysis of error propagation in the PII propensity scale was calculated in Python, where the error is the standard deviation in the difference in ΔG_{app} from the fit K_{app}.

All CD scans were performed on a Jasco J-720 spectropolarimeter. Sos peptide samples were prepared by diluting ITC ligand solutions (with identical buffer) to concentrations suitable for CD, ~0.1 mg/mL. As such, the buffer conditions were identical to ITC (20 m*M* sodium phosphate, 200 m*M* sodium chloride, pH 7.5). Spectra were measured at 298 K from 200 to 250 nm, at a scan rate of 10 s/nm. Data were collected in nanometer increments and represent an average of three scans.

A detailed derivation is provided in the Supporting Information to explain how PII propensities are determined.

Calculation of PII propensity in sequences and in silico *evolution of PII propensity*

Several protein sequence datasets[39,40,52] were employed for our analysis. Several protein sequence datasets were employed for the analysis of the PII content of the proteome, including a nonredundant set of human protein sequences extracted from the PDB[39] consisting of proteins from each SCOP family,[57] an ID protein dataset DisProt 5.5,[40] and the complete proteomes of six eukaryotes—*H. sapiens* (human), *M. musculus* (mouse), *D. melanogaster* (fly), *C. elegans* (worm), *A. thaliana* (plant), *S. cerevisiae* (yeast) obtained from the Integr8 project.[52] Algorithms for calculating the PII propensities of amino acid sequences were written in C++ and Python, with additional data processing in Perl, the R Project, and Microsoft Excel. The PII bias at a specific position was computed as an average of the PII propensities for a given window. The process is customizable- varying propensity scales, inclusion/exclusion of amino acids, window size (1–100) have been tested (Fig. 2 and Supporting Information Fig. 2). The calculations can be performed ignoring the contributions of specific amino acids (such as proline, for example) and with any window size. Statistical difference between PII propensity distributions were assessed by a t-test. Calculations shown in Figures 2 and 5 were performed over a window size of 32 residues, conservatively assuming 60% PII propensity for proline. Each distribution is comprised of over 40,000 data points. In Figure 4(A–C), a window size of 10 residues was used. Statistical significance of correlations shown in Figure 4(F) were assessed by mathematical standards.[58]

An algorithm was developed to compare the PII content of a biological "parent" sequence to those created by random substitution (maintaining background amino acid frequencies of the datasets) or by substitution that conserves physico-chemical properties, BLOSUM62[45] or PAM[59] substitution matrices. To quantify the effect of mutation on PII content of a sequence, the average PII propensities of "parent" and the *in silico* evolved "daughter" sequences were calculated as described above. Substitution within the algorithm is completely adjustable, and can involve any part of the sequence, maintaining any arbitrary level of identity to the parent sequence, and with any input substitution frequencies. The means and standard deviations of the PII distributions for *in silico* evolved "daughter" sequences (n = 10,000) were computed from which the z-score of the average PII propensity of the "parent" sequences was calculated. The z-score was converted into a p-value, the logs of which are plotted in Figure 6. In all cases except the PII distribution of "daughter" high PII, BLOSUM62[45] sequences, the distributions were normal. Code and scripts for shuffling tests, *in silico* evolution, and analyzing the PII content of the sequences were written in Python.

Gene ontology of high PII proteins

The database for annotation, visualization, and integrated discovery (DAVID) was identified enriched features and functions of the top 1% of high PII proteins in eukaryotic proteomes obtained from Integr8.[52] To assess GO term enrichment in the top 1% of proteins selected for GO analysis (n = 200), a one-tailed Fisher exact test was used,[60] and P-values of enrichment reported by DAVID[61,62] were normalized to P-values that could be obtained by submission of random protein datasets (10^{-3}). In Figure 7(B), reported is the number of GO terms returned with significance ($P \leq 10^{-6}$), with "phosphoprotein" enriched ($P \leq 10^{-8}$) in all three species. The exact normalized p-values for "phosphoprotein" enrichment in *H. sapiens*, *M. musculus*, and *S. cerevisiae* [Fig.

4(B)] were $P = 7.6 \times 10^{-11}$, $P = 5.4 \times 10^{-8}$, and $P = 8.7 \times 10^{-14}$, respectively. Code and scripts for mining the proteomes and analyzing the sequences were written in Python and Perl. DAVID output was processed manually in Microsoft excel. Calculations for Figures 7–9 were performed over a window size of 50 residues, assuming 60% PII propensity for proline. Each distribution is comprised of over 10,000 data points. Gene lists of proteins extracted from the proteome based on their PII propensities were nonredundant. High PII proteins (top 1% for GO analysis) had the longest continuous segments of high PII bias. Shown in Figure 7(D) is the log of the p-value for "phosphoprotein" divided by log 10^{-3} (noise). In Figure 8(C), the number of phosphorylation sites of each type (SER, THR, or TYR) is normalized by the number of peptides in each PII bin from the *H. sapiens* proteome. Phosphoproteins were classified as low PII (PII% < 31%, one standard deviation below the mean) or high PII (PII% >38%, one standard deviation above the mean) based upon the sequence PII context of their experimentally validated phosphorylation sites. The sizes of the pie pieces in Figure 9 correspond to the relative statistical enrichment of the GO term obtained from DAVID for phosphoserine, phosphothreonine, and phosphotyrosine containing proteins that were then grouped in the pie representation to show relative enrichment. GO terms grouped as "Other" (Fig. 9) were individually statistically enriched ($P < 0.05$), yet relative to other GO terms comprised less than 1% (individually) of the pie and were grouped for clarity.

Acknowledgments

The authors acknowledge James O. Wrabl and Jiin-Yu Chen for reviewing the manuscript and George D. Rose for insightful discussions. WAE, TPS, and VJH designed the research. WAE, TPS, and AJC performed the work and analyzed the data. WAE and VJH wrote the paper.

References

1. Ward J, Sodhi J, McGuffin L, Buxton B, Jones D (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J Mol Biol 337:635–645.
2. Brown C, Johnson A, Daughdrill G (2010) Comparing models of evolution for ordered and disordered proteins. Mol Biol Evol 27:609–621.
3. Oldfield C, Cheng Y, Cortese M, Brown C, Uversky V, Dunker A (2005) Comparing and combining predictors of mostly disordered proteins. Biochemistry 44: 1989–2000.
4. Vacic V, Oldfield C, Mohan A, Radivojac P, Cortese M, Uversky V, Dunker A (2007) Characterization of molecular recognition feautures, MoRFs, and their binding partners. J Proteome Res 6:2351–2366.
5. Disfani F, Hsu W, Mizianty M, Oldfield C, Xue B, Dunker A, Uversky V, Kurgan L (2012) MoRFpred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins. Bioinformatics 28:i75–i83.
6. Tiffany M, Krimm S (1968) Circular dichroism of poly-L-proline in an unordered conformation. Biopolymers 6: 1767–1770.
7. Tiffany M, Krimm S (1973) Extended conformations of polypeptides and proteins in urea and guanidine hydrochloride. Biopolymers 12:575–587.
8. Krimm S, Mark J (1968) Conformations of polypeptides with ionized side chains of equal length. Proc Natl Acad Sci USA 60:1122–1129.
9. Tiffany M, Krimm S (1968) New chain conformations of poly(glutamic acid) and polylysine. Biopolymers 6: 1379–1382.
10. Chellgren B, Miller A, Creamer T (2006) Evidence of polyproline II helical structure in short polyglutamine tracts. J Mol Biol 361:362–371.
11. Mao A, Crick S, Vitalis A, Chicoine C, Pappu R (2010) Net charge per residue modulates conformational ensembles of intrinsically disordered proteins. Proc Natl Acad Sci USA 107:8183–8188.
12. Williamson M (1994) The structure and function of proline-rich regions in proteins. Biochem J 297:249–260.
13. Kay B, Williamson M, Sudol M (2000) The importance of being proline: the interaction of proline-rich motifs in signaling proteins with their cognate domains. FASEB J 14:231–241.
14. Pawson T, Nash P (2003) Assembly of cell regulatory systems through protein interaction domains. Sci Signal 300:445–452.
15. Zarrinpar A, Bhattacharyya R, Lim W (2003) The structure and function of proline recognition domains. Sci Signal 179:1–10.
16. Rubin G, Yandell M, Wortman J, Miklos G, Nelson C, Hariharan I, Fortini M, Li P, Apweiler R, Fleischmann W, Cherry J, Henikoff S, Skupski M, Misra S, Ashburner M, Birney E, Boguski M, Brody T, Brokstein P, Celniker S, Chervitz S, Coates D, Cravchik A, Gabrielan A, Galle R, Gelbart W, George R, Goldstein L, Gong F, Guan P, Harris N, Hay B, Hoskins R, Li J, Li Z, Hynes R, Jones S, Kuehl P, Lemaitre B, Littleton J, Morrison D, Mungall C, O'Farrell P, Pickeral O, Shue C, Vosshall L, Zhang J, Zhao Q, Zheng X, Zhong F, Zhong W, Gibbs R, Venter J, Adams M, Lewis S (2000) Comparative genomics of the eukaryotes. Science 287: 2204–2215.
17. Rucker A, Pager C, Campbell M, Qualls J, Creamer T (2003) Host-guest scale of left-handed polyproline II helix formation. Proteins 53:68–75.
18. Shi Z, Chen K, Liu Z, Ng A, Bracken W, Kallenbach N (2005) Polyproline II propensities from GGXGG peptides reveal an anticorrelation with B-sheet scales. Proc Natl Acad Sci USA 102:17964–17968.
19. Grdadolnik J, Mohacek-Grosev V, Baldwin R, Avbelji F (2011) Populations of the three major backbone conformations in 19 amino acid dipeptides. Proc Natl Acad Sci USA 108:1794–1798.
20. Oh K, Lee K, Park E, Jung Y, Hwang G, Cho M (2012) A comprehensive library of blocked dipeptides reveals intrinsic backbone conformational propensities of unfolded proteins. Proteins 80:977–990.
21. Brown A, Zondlo N (2012) A propensity scale for type II polyproline helices (PPII): aromatic amino acids in proline-rich sequences strongly disfavor PPII due to proline-aromatic interactions. Biochemistry 51: 5041–5051.
22. Tran H, Wang X, Pappu R (2005) Reconciling observations of sequence-specific conformational propensities

with the generic polymeric behavior of denatured proteins. Biochemistry 44:11369–11380.

23. Fleming P, Fitzkee N, Mezei M, Srinivasan R, Rose G (2005) A novel method reveals that solvent water favors polyproline II over B-strand conformation in peptides and unfolded proteins: conditional hydrophobic accessible surface area (CHASA). Protein Sci 14: 111–118.
24. Beck D, Alonso D, Inoyama D, Daggett V (2008) The intrinsic conformational propensities of the 20 naturally ocurring amino acids and reflection of these propensities in proteins. Proc Natl Acad Sci USA 105: 12259–12264.
25. Moradi M, Babin V, Sagui C, Roland C (2011) A statistical analysis of the PPII propensity of amino acid guests in proline-rich peptides. Biophys J 100: 1083–1093.
26. Ferreon J, Hilser V (2003) The effect of the polyproline II (PPII) conformation on the denatured state entropy. Protein Sci 12:447–457.
27. Ferreon J, Hilser V (2004) Thermodynamics of binding to SH3 domains: the energetic impact of polyproline II (PII) helix formation. Biochemistry 43:7787–7797.
28. Hamburger J, Ferreon J, Whitten S, Hilser V (2004) Thermodynamic mechanism and consequences of the polyproline II (PII) structural bias in the denatured states of proteins. Biochemistry 43:9790–9799.
29. Lim W, Richards F, Fox R (1994) Structural determinants of peptide-binding orientation and of sequence specificity in SH3 domains. Nature 372:375–379.
30. Dasgupta B, Chakrabarti P, Basu G (2007) Enhanced stability of cis Pro-Pro peptide bond in the Pro-Pro-Phe sequence motif. FEBS Lett 581:4529–4532.
31. Ganguly H, Majumder B, Chattopadhyay S, Chakrabarti P, Basu G (2012) Direct evidence for CH-pi interaction mediated stabilization of Pro-cisPro bond in peptides with Pro-Pro-aromatic motifs. J Am Chem Soc 134:4661–4669.
32. Elam W, Schrank T, Hilser V, Experimental and computation studies of polyproline II propensity. In: Schweitzer-StennerR, Ed. (2012) Protein and peptide folding, misfolding, and non-folding. Hoboken, NJ: Wiley, pp159–185.
33. Vila J, Baldoni H, Ripoll D, Ghosh A, Scheraga H (2004) Polyproline II helix conformation in a proline-rich environment: a theoretical study. Biophys J 6: 731–742.
34. Whitten S, Yang H, Fox R, Hilser V (2008) Exploring the impact of polyproline II (PII) conformational bias on the binding of peptides to the Sem-5 SH3 domain. Protein Sci 17:1200–1211.
35. Bielska A, Zondlo N (2006) Hyperphosphorylation of tau induces local polyproline II helix. Biochemistry 45: 5527–5537.
36. Huber T, Grama L, Hetenyi C, Schay G, Fulop L, Penke B, Kellermayer M (2012) Conformational dynamics of titin PEVK explored with FRET spectroscopy. Biophys J 103:1480–1489.
37. Kohler S, Weber A, Howard S, Welte W, Drescher M (2010) The proline-rich domain of TonB possesses an extended polyproline II-like conformation of sufficient length to span the periplasm of gram-negative bacteria. Protein Sci 19:625–630.
38. Clantin B, Delattre A, Rucktooa P, Saint N, Meli A, Locht C, Jacob-Dubuisson F, Villeret V (2007) Structure of the membrane protein FhaC: a member of the Omp85-TpsB transporter superfamily. Science 317: 957–961.
39. Berman H, Westbrook J, Feng Z, Gilliland G, Bhat T, Weissig H, Shindylalov I, Bourne P (2000) The protein data bank. Nucl Acids Res 28:235–242.
40. Sickmeier M, Hamilton J, LeGall T, Vacic V, Cortese M, Tantos A, Szabo B, Tompa P, Chen J, Uversky V, Obradovic Z, Dunker A (2007) DisProt: the database of disordered proteins. Nucl Acids Res 35:D786–D793.
41. Avbelj F, Baldwin R (2004) Origin of the neighboring residue effect on peptide conformation. Proc Natl Acad Sci USA 101:10967–10972.
42. Avbelj F, Golic-Grdadolnik S, Grdadolnik J, Baldwin R (2006) Intrinsic backbone preferences are fully present in blocked amino acids. Proc Natl Acad Sci USA 103: 1272–1277.
43. Campen A, Williams R, Brown C, Meng J, Uversky V, Dunker A (2008) TOP-IDP scale: a new amino acid scale measuring propensity of intrinsic disorder. Protein Pept Lett 15:956–963.
44. Afek A, Shakhnovich E, Lukatsky D (2011) Multi-scale sequence correlations increase proteome structural disorder and promiscuity. J Mol Biol 409:439–449.
45. Henikoff S, Henikoff J (1992) Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 89:10915–10919.
46. Schaefer C, Schlessinger A, Rost B (2010) Protein secondary structure appears to be robust under in silico evolution while protein disorder appears not to be. Bioinformatics 26:625–631.
47. Pauling L, Corey R (1951) The structure of fibrous proteins of the collagen-gelatin group. Proc Natl Acad Sci USA 37:272–281.
48. Liu J, Peruma N, Oldfield C, Su E, Uversky V, Dunker A (2006) Intrinsic disorder in transcription factors. Biochemistry 45:6873–6888.
49. Iakoucheva L, Radivojac P, Brown C, O'Connor T, Sikes J, Obradovic Z, Dunker A (2004) The importance of intrinsic disorder for protein phosphorylation. Nucl Acids Res 32:1037–1049.
50. D'Aquino J, Gomez J, Hilser V, Lee K, Amzel L, Freire E (1996) The magnitude of the backbone conformational entropy change in protein folding. Proteins 25: 143–156.
51. Kim S, Jung Y, Hwang G, Han H, Cho M (2011) Phosphorylation alters backbone conformational preferences of serine and threonine peptides. Proteins 79: 3155–3165.
52. Kersey P, Bower L, Morris L, Horne A, Petryszak R, Kanz C, Kanapin A, Das U, Michoud K, Phan I, Gattiker A, Kulikova T, Faruque N, Duggan K, Mclaren P, Reimholz B, Duret L, Penel S, Reuter I, Apweiler R (2005) Integr8 and genome reviews: integrated views of complete genomes and proteomes. Nucl Acids Res 33: D297–D302.
53. Farriol-Mathis N, Garavelli J, Boeckmann B, Duvaud S, Gasteiger E, Gateau A, Veuthey A, Bairoch A (2004) Annotation of post-translational modifications in the Swiss-Prot knowledge base. Proteomics 4:1537–1550.
54. Chen S, Chen F, Li W (2010) Phosphorylated and non-phosphorylated serine and threonine residues evolve at different rates in mammals. Mol Biol Evol 27: 2548–2554.
55. Wiseman T, Williston S, Brandts J, Lin L (1989) Rapid measurement of binding constants and heats of binding using a new titration calorimeter. Analyt Biochemy 179:131–137.
56. Edelhoch H (1967) Spectroscopic determination of tryptophan and tyrosine in proteins. Biochemistry 6: 1948–1954.

57. Murzin A, Brenner S, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247:536–540.

58. Zar J (1972) Significance testing of the Spearman rank correlation coefficient. J Am Stat Assoc 67:578–580.

59. Dayhoff M, Schwartz R, Orcutt B (1978) A model of evolutionary change in proteins. Atlas Protein Sequence Struct Chapter 22:345–352.

60. Hosack D, Dennis G, Sherman B, Lane H, Lempicki R (2003) Identifying biological themes within lists of genes with EASE. Genome Biol 4:R70.71–R70.78.

61. Dennis G, Sherman B, Hosack D, Yang J, Gao W, Lane H, Lempicki R (2003) DAVID: database for annotation, visualization, and integrated discovery. Genome Biol 4: 3.

62. Huang D, Sherman B, Lempicki R (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4:44–57.